chapter 1

Generative Grammar

0. Preliminaries

Although we use it every day, and although we all have strong opinions about its proper form and appropriate use, we rarely stop to think about the wonder of language. So-called language "experts" like William Safire tell us about the misuse of *hopefully* or lecture us about the origins of the word *boondoggle*, but surprisingly, they never get at the true wonder of language: how it actually works. Think about it for a minute; you are reading this and understanding it but you have no conscious knowledge of how you are doing it. The study of this mystery is the science of linguistics. This book is about one aspect of how language works – how sentences are structured: ***syntax***.

Language is a psychological or cognitive property of humans. That is, there is some set of neurons in my head firing madly away that allows me to sit here and produce this set of letters, and there is some other set of neurons in your head firing away that allows you to translate these squiggles into coherent ideas and thoughts. There are several subsystems at work here. If you were listening to me speak, I would be producing sound waves with my vocal cords and articulating particular speech sounds with my tongue, lips, and vocal cords. On the other end of things you'd be hearing those sound waves and translating them into speech sounds using your auditory apparatus. The study of the acoustics and articulation of speech is called ***phonetics***. Once you've translated the waves of sound into mental representations of speech sounds, you analyze them into syllables and

pattern them appropriately. For example, speakers of English know that the made-up word *bluve* is a possible word of English, but the word *bnuck* is not. This is part of the science called ***phonology***. Then you take these groups of sounds and organize them into meaningful units (called morphemes) and words. For example, the word *dancer* is made up of two meaningful bits: *dance* and the suffix -*er*. The study of this level of Language is called ***morphology***. Next you organize the words into phrases and sentences. ***Syntax*** is the cover term for studies of this level of Language. Finally, you take the sentences and phrases you hear and translate them into thoughts and ideas. This last step is what we refer to as the ***semantic*** level of Language.

Syntax, then, studies the level of Language that lies between words and the meaning of utterances: sentences. It is the level that mediates between sounds that someone produces (organized into words) and what they intended to say.

Perhaps one of the truly amazing aspects of the study of Language is not the origins of the word *demerit,* or how to properly punctuate a quote inside parentheses, or how kids have, like, destroyed the English language, eh? Instead it's the question of how we subconsciously get from sounds to meaning. This is the study of syntax.

Language vs. language

When I utter the term *language*, most people immediately think of some particular language such as English, French, or KiSwahili. But this is not the way linguists use the term; when linguists talk about ***Language*** (or i-language), they are generally talking about the *ability* of humans to speak any (particular) language. Some people (most notably Noam Chomsky) also call this the ***Human Language Capacity***. Language (written with a capital L) is the part of the mind or brain that allows you to speak, whereas ***language*** (with a lower case l) (also known as e-language) is an instantiation of this ability (like French or English). In this book we'll be using language as our primary data, but we'll be trying to come up with a model of Language.

1. Syntax as a Cognitive Science

Cognitive science is a cover term for a group of disciplines that all aim for the same goal: describing and explaining human beings' ability to think (or more particularly, to think about abstract notions like subatomic particles,

the possibility of life on other planets or even how many angels can fit on the head of a pin, etc.). One thing that distinguishes us from other animals, even relatively smart ones like chimps and elephants, is our ability to use productive, combinatory Language. Language plays an important role in how we think about abstract notions, or, at the very least, Language appears to be structured in such a way that it allows us to express abstract notions.[1] The discipline of linguistics, along with psychology, philosophy, and computer science, thus forms an important subdiscipline within cognitive science. Sentences are how we get at expressing abstract thought processes, so the study of syntax is an important foundation stone for understanding how we communicate and interact with each other as humans.

2. MODELING SYNTAX

The dominant theory of syntax is due to Noam Chomsky and his colleagues, starting in the mid 1950s and continuing to this day. This theory, which has had many different names through its development (Transformational Grammar (TG), Transformational Generative Grammar, Standard Theory, Extended Standard Theory, Government and Binding Theory (GB), Principles and Parameters approach (P&P) and Minimalism (MP)), is often given the blanket name ***Generative Grammar***. A number of alternate theories of syntax have also branched off of this research program; these include Lexical-Functional Grammar (LFG) and Head-Driven Phrase Structure Grammar (HPSG). These are also considered part of generative grammar; but we won't cover them extensively in this book, except in chapters 16 and 17. The particular version of generative grammar that we will mostly look at here is roughly the ***Principles and Parameters*** approach, although we will occasional stray from this into the more recent version called ***Minimalism***.

The underlying thesis of generative grammar is that sentences are generated by a subconscious set of procedures (like computer programs). These procedures are part of our minds (or of our cognitive abilities if you prefer). The goal of syntactic theory is to model these procedures. In other words, we are trying to figure out what we subconsciously know about the syntax of our language.

In generative grammar, the means for modeling these procedures is through a set of formal grammatical ***rules***. Note that these rules are nothing

[1] Whether language constrains what abstract things we can think about (this idea is called the Sapir-Whorf hypothesis) is a matter of great debate and one that lies outside the domain of syntax per se.

like the rules of grammar you might have learned in school. These rules don't tell you how to properly punctuate a sentence or not to split an infinitive. Instead, they tell you the order in which to put your words (in English, for example, we put the subject of a sentence before its verb; this is the kind of information encoded in generative rules). These rules are thought to generate the sentences of a language, hence the name *generative* grammar. You can think of these rules as being like the command lines in a computer program. They tell you step by step how to put together words into a sentence. We'll look at precise examples of these rules in the next chapter. But before we can get into the nitty-gritty of sentence structure, let's look at some of the underlying assumptions of generative grammar.

Noam Chomsky

Avram Noam Chomsky was born on the 7th of December 1928, in Philadelphia. His father was a Hebrew grammarian and his mother a teacher. Chomsky got his Ph.D. from the University of Pennsylvania, where he studied linguistics under Zellig Harris. He took a position in machine translation and language teaching at the Massachusetts Institute of Technology. Eventually his ideas about the structure of language transformed the field of linguistics. Reviled by some and admired by others, Chomsky's ideas have laid the groundwork for the discipline of linguistics, and have been very influential in computer science, and philosophy.

Chomsky is also one of the leading intellectuals in the anarchist socialist movement. His political writings about the media and political injustice have profoundly influenced many. Chomsky is among the most quoted authors in the world (among the top ten and the only living person on the list).

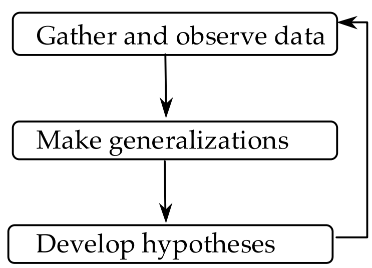
3. Syntax as Science – the Scientific Method

To many people the study of language properly belongs in the domain of the humanities. That is, the study of language is all about the beauty of its usage in fine (and not so fine) literature. However, there is no particular reason, other than our biases, that the study of language should be confined to a humanistic approach. It is also possible to approach the study of language from a scientific perspective; this is the domain of linguistics. People who study literature often accuse linguists of abstracting away from the richness of good prose and obscuring the beauty of language. Nothing could be further from the truth. Most linguists, including the present author,

enjoy nothing more than reading a finely crafted piece of fiction, and many linguists often study, as a sideline, the more humanistic aspects of language. This doesn't mean, however, that one can't appreciate and study the formal properties (or rules) of language and do it from a scientific perspective. The two approaches to language study are both valid; they complement each other; and neither takes away from the other.

Science is perhaps one of the most poorly defined words of the English language. We regularly talk of scientists as people who study bacteria, particle physics, and the formation of chemical compounds, but ask your average Joe or Jill on the street what science means, and you'll be hard pressed to get a decent definition. Science refers to a particular methodology for study: the scientific method. The scientific method dates backs to the ancient Greeks, such as Aristotle, Euclid, and Archimedes. The method involves observing some data, making some generalizations about patterns in the data, developing hypotheses that account for these generalizations, and testing the hypotheses against more data. Finally, the hypotheses are revised to account for any new data and then tested again. A flow chart showing the method is given in (1):

1)



In syntax, we apply this methodology to sentence structure. Syntacticians start[2] by observing data about the language they are studying, then they make generalizations about patterns in the data (e.g., in simple English declarative sentences, the subject precedes the verb). They then generate a hypothesis – preferably one that makes predictions – and test the hypothesis against more syntactic data, and if necessary go back and re-evaluate their hypotheses.

---

[2] This is a bit of an oversimplification. We really have a "chicken and the egg" problem here. You can't know what data to study unless you have a hypothesis about what is important, and you can't have a hypothesis unless you have some basic understanding of the data. Fortunately, as working syntacticians this philosophical conundrum is often irrelevant, as we can just jump feet-first into both the hypothesis-forming and data-analysis at the same time.

Hypotheses are only useful to the extent that they make ***predictions***. A hypothesis that makes no predictions (or worse yet, predicts everything) is useless from a scientific perspective. In particular, the hypothesis must be ***falsifiable***. That is we must, in principle, be able to look for some data, which if true, show that the hypothesis is wrong. This means that we are often looking for the cases where our hypotheses predict that a sentence will be grammatical (and it is not), or the cases where they predict that the sentence will be ungrammatical (but it is).

In syntax, hypotheses are called ***rules,*** and the group of hypotheses that describe a language's syntax is called a ***grammar***.

Do Rules Really Exist?

Generative grammar claims to be a theory of cognitive psychology, so the natural question to ask at this point is whether formal rules really exist in the brain/minds of speakers. After all, a brain is a mass of neurons firing away, how can formal mathematical rules exist up there? Remember, however, that we are attempting to *model* Language, we aren't trying to describe Language exactly. This question confuses two disciplines: psychology and neurology. Psychology is concerned with the mind, which represents the output and the abstract organization of the brain. Neurology is concerned with the actual firing of the neurons and the physiology of the brain. Generative grammar doesn't try to be a theory of neurology. Instead it is a model of the psychology of Language. Obviously, the rules don't exist, per se in our brains, but they do represent the external behavior of the mind. For more discussion of this issue, look at the readings in the further reading section of this chapter.

The term *grammar* strikes terror into the hearts of many people. But you should note that there are two ways to go about writing grammatical rules. One is to tell people how they should speak (this is of course the domain of English teachers and copy-editors); we call these kinds of rule ***prescriptive rules*** (as they prescribe how people should speak according to some standard). Some examples of prescriptive rules include "never end a sentence with a preposition," "use *whom* not *who*," "don't split infinitives." These rules tell us how we are supposed to use our language. The other approach is to write rules that describe how people *actually* speak, whether or not they are speaking "correctly." These are called ***descriptive rules***. Consider for a moment the approach we're taking in this book; which of the two types (descriptive or prescriptive) is more scientific? Which kind of rule is more likely to give us insight into how the mind uses Language? For these

reasons, we focus on descriptive rules. This doesn't mean that prescriptive rules aren't important (in fact, in the problem sets section of this chapter you are asked to critically examine the question of descriptive vs. prescriptive rules), but for our purposes descriptive rules are more important. For an interesting discussion of the prescriptive/descriptive debate, see Pinker's 1995 book: *The Language Instinct*.

You now have enough information to answer General Problem Set 1

3.1 An Example of the Scientific Method as Applied to Syntax

Let's turn now to a real world application of the scientific method to some language data. The following data concern the form of a specific kind of noun, called an ***anaphor*** (plural: ***anaphors***, the phenomenon is called ***anaphora***). These are the nouns that end with *-self* (e.g., *himself, herself, itself,* etc.). In chapter 5, we look at the distribution of anaphora in detail; here we'll only consider one superficial aspect of them. In the following sentences, as is standard in the syntactic literature, a sentence that isn't well-formed is marked with an ***asterisk*** (*) before it. For these sentences assume that *Bill* is male and *Sally* is female.

2) a) Bill kissed himself.
   b) *Bill kissed herself.
   c) Sally kissed herself.
   d) *Sally kissed himself.
   e) *Kiss himself.

To the unskilled eye, the ill-formed sentences in (2b and d) just look silly. It is obvious that Bill can't kiss herself, because Bill is male. However, no matter how matter-of-factly obvious this is, it is part of a bigger generalization about the distribution of anaphors. In particular, the generalization we can draw about the sentences in (2) is that an anaphor must agree in ***gender*** with the noun it refers to (its ***antecedent***). So in (2a and b) we see that the anaphor must agree in gender with *Bill,* its antecedent. The anaphor must take the masculine form *himself*. The situation in (2c and d) is the same; the anaphor must take the form *herself* so that it agrees in gender with the feminine *Sally*. Note further that a sentence like (2e) shows us that anaphors must have an antecedent. An anaphor without an antecedent is unacceptable. A plausible hypothesis (or rule) given the data in (2), then, is stated in (3):

3) An anaphor must (i) have an antecedent and (ii) agree in gender (masculine, feminine, or neuter) with that antecedent.

The next step in the scientific method is to test this hypothesis against more data. Consider the additional data in (4):

4) a) The robot kissed itself.
   b) She knocked herself on the head with a zucchini.
   c) *She knocked himself on the head with a zucchini.
   d) The snake flattened itself against the rock.
   e) ?The snake flattened himself/herself against the rock.
   f) The Joneses think themselves the best family on the block.
   g) *The Joneses think himself the most wealthy guy on the block.
   h) Gary and Kevin ran themselves into exhaustion.
   i) *Gary and Kevin ran himself into exhaustion.

Sentences (4a, b, and c) are all consistent with our hypothesis that anaphors must agree in gender with their antecedents, which at least confirms that the hypothesis is on the right track. What about the data in (4d and e)? It appears as if any gender is compatible with the antecedent *the snake*. This appears, on the surface, to be a contradiction to our hypothesis. Think about these examples a little more closely, however. Whether sentence (4e) is well-formed or not depends upon your assumptions about the gender of the snake. If you assume (or know) the snake to be male, then *The snake flattened himself against the rock* is perfectly well-formed. But under the same assumption, the sentence *The snake flattened herself against the rock* seems very odd indeed, although it is fine if you assume the snake is female. So it appears as if this example also meets the generalization in (3); the vagueness about its well-formedness has to do with the fact that we are rarely sure what gender a snake is and not with the actual structure of the sentence.

Now, look at the sentences in (4f–i); note that the ill-formedness of (g) and (i) is not predicted by our generalization. In fact, our generalization predicts that sentence (4i) should be perfectly grammatical, since *himself* agrees in gender (masculine) with its antecedents *Gary* and *Kevin*. Yet there is clearly something wrong with this sentence. The hypothesis needs revision. It appears as if the anaphor must agree in gender and ***number*** with the antecedent. Number refers to the quantity of individuals involved in the sentence; English primarily distinguishes singular number from plural number. (5) reflects our revised hypothesis.

5) An anaphor must agree in gender and number with its antecedent.

If there is more than one person or object mentioned in the antecedent, then the anaphor must be plural (i.e., *themselves*).

Testing this against more data, we can see that this partially makes the right predictions (6a), but it doesn't properly predict the grammaticality of sentences (6b–e):

6) a) People from Tucson think very highly of themselves.
   b) *I gave yourself the bucket of ice cream.
   c) I gave myself the bucket of ice cream.
   d) *She hit myself with a hammer.
   e) She hit herself with a hammer.

Even more revision is in order. The phenomenon seen in (6b–e) revolves around a grammatical distinction called ***person***. Person refers to the perspective of the speaker with respect to the other participants in the speech act. First person refers to the speaker. Second person refers to the listener. Third person refers to people being discussed that aren't participating in the conversation. Here are the English pronouns associated with each person: (***Nominative*** refers to the ***case*** form the pronouns take when in subject position like *I* in "*I* love peanut butter;" ***accusative*** refers to the form they take when in object positions like *me* in "John loves *me*." We will look at case in much more detail in chapter 9, so don't worry if you don't understand it right now.)

7)

	Nominative		Accusative		Anaphoric	
	Singular	Plural	Singular	Plural	Singular	Plural
1	I	we	me	us	myself	ourselves
2	you	you	you	you	yourself	yourselves
3 masc	he		him		himself	
3 fem	she	they	her	them	herself	themselves
3 neut	it		it		itself	

As you can see from this chart, the form of the anaphor seems also to agree in person with its antecedent. So once again we revise our hypothesis (rule):

8) An anaphor must agree in person, gender and number with its antecedent.

With this hypothesis, we have a straightforward statement of the distribution of this noun type, derived using the scientific method. In the problem sets below, and in chapter 5, you'll have an opportunity to revise the rule in (8) with even more data.

You now have enough information to try Challenge Problem Sets 1 & 2

3.2 Sources of Data

If we are going to apply the scientific method to syntax, it is important to consider the sources of data. One obvious source is in collections of either spoken or written texts. Such data are called ***corpora*** (singular: ***corpus***). There are many corpora available, including some searchable through the World Wide Web. For languages without a literary tradition or ones spoken by a small minority, it is often necessary for the linguist to go and gather data and compile a corpus in the field. In the early part of this century, this was the primary occupation of linguists, and it is proudly carried on today by many researchers.

While corpora are unquestionably invaluable sources of data, they are only a partial representation of what goes on in the mind. More particularly, corpora often contain instances of only grammatical (or more precisely well-formed) sentences (sentences that sound "OK" to a native speaker). For example, the online New York Times contains very few ungrammatical sentences. Even corpora of naturalistic speech complete with the errors every speaker makes don't necessarily contain the data we need to test the falsifiable predictions of our hypotheses.

You might think that what's in a corpus would be enough for a linguist to do her job. But corpora are just not enough: there is no way of knowing whether a corpus has *all* possible forms of grammatical sentences. In fact, as we will see in the next chapter, due to the productive nature of language, a corpus could *never* contain all the grammatical forms of a language, nor could it even contain a representative sample. To really get at what we know about our languages (remember syntax is a cognitive science), we have to know what sentences are *not* well-formed. That is, in order to know the range of what are acceptable sentences of English, Italian or Igbo, we *first* have to know what are *not* acceptable sentences in English, Italian or Igbo. This kind of negative information is very rarely available in corpora, which mostly provide grammatical, or well-formed, sentences.

Consider the following sentence:

9) *Who do you wonder what bought?

For most speakers of English, this sentence borders on word salad – it is not a good sentence of English. How do you know that? Were you ever taught in school that you can't say sentences like (9)? Has anyone ever uttered this sentence in your presence before? I seriously doubt it. The fact that a sentence like (9) sounds strange, but similar sentences like (10a and b) *do* sound OK is not reflected anywhere in a corpus:

10) a) Who do you think bought the bread machine?
    b) I wonder what Fiona bought.

Instead we have to rely on our knowledge of our native language (or on the knowledge of a native speaker consultant for languages that we don't speak natively). Notice that this is *not* conscious knowledge. I doubt there are many native speakers of English that could tell you why sentence (9) is terrible, but most can tell you that it is. This is subconscious knowledge. The trick is to get at and describe this subconscious knowledge.

The psychological experiment used to get this subconscious kind of knowledge is called the ***grammaticality judgment task***. The judgment task involves asking a native speaker to read a sentence, and judge whether it is well-formed (grammatical), marginally well-formed, or ill-formed (unacceptable or ungrammatical).

Judgments as Science?

Many linguists refer to the grammaticality judgment task as "drawing upon our native speaker intuitions." The word "intuition" here is slightly misleading. The last thing that pops into our heads when we hear the term "intuition" is science. Generative grammar has been severely criticized by many for relying on "unscientific" intuitions. But this is based primarily on a misunderstanding of the term. To the lay person, the term "intuition" brings to mind guesses and luck. This usage of the term is certainly standard. When a generative grammarian refers to "intuition" however, she is using the term to mean "tapping into our subconscious knowledge." The term "intuition" may have been badly chosen, but in this circumstance it refers to a real psychological effect. Intuition (as a grammaticality judgment) has an entirely scientific basis. It is replicable under strictly controlled experimental conditions (these conditions are rarely applied, but the validity of the task is well established). Other disciplines also use intuitions or judgment tasks. For example, within the study of vision, it has been determined that people can accurately judge differences in light intensity, drawing upon their subconscious knowledge (Bard et al. 1996). To avoid the negative associations with the term intuition, we will use the term ***judgment*** instead.

There are actually several different kinds of grammaticality judgments. Both of the following sentences are ill-formed, but for different reasons:

11) a) #The toothbrush is pregnant.
    b) *Toothbrush the is blue.

Sentence (11a) sounds bizarre (cf. *the toothbrush is blue*) because we know that toothbrushes (except in the world of fantasy/science fiction or poetry) cannot be pregnant. The meaning of the sentence is strange, but the form is OK. We call this ***semantic ill-formedness*** and mark the sentence with a #. By contrast, we can glean the meaning of sentence (11b); it seems semantically reasonable (toothbrushes can be blue), but it is ill-formed from a structural point of view. That is, the determiner *the* is in the wrong place in the sentence. This is a ***syntactically ill-formed*** sentence. A native speaker of English will judge both these sentences as ill-formed, but for very different reasons. In this text, we will be concerned primarily with syntactic well-formedness.

You now have enough information to answer General Problem Set 2

---

4. WHERE DO THE RULES COME FROM?

In this chapter we've been talking about our subconscious knowledge of syntactic rules, but we haven't dealt with how we get this knowledge. This is sort of a side issue, but it may affect the shape of our theory. If we know how children acquire their rules, then we are in a better position for a proper formalization of them. The way by which children develop knowledge is an important question in cognitive science. The theory of generative grammar makes some very specific (and very surprising) claims about this.

4.1 Learning vs. Acquisition

One of the most common misconceptions about Language is the idea that children and adults "learn" languages. Recall that the basic kind of knowledge we are talking about here is subconscious knowledge. When producing a sentence you don't consciously think about where to put the subject, where to put the verb, etc. Your subconscious language faculty does that for you. Cognitive scientists make a distinction in how we get conscious and subconscious knowledge. Conscious knowledge (like the rules of algebra, syntactic theory, principles of organic chemistry, or how to take apart a carburetor) is ***learned***. Subconscious knowledge, like how to speak or the ability to visually identify discrete objects, is ***acquired***. In part, this explains why classes in the formal grammar of a foreign language often fail abysmally to train people to speak those languages. By contrast, being immersed in an environment where you can subconsciously acquire a language is much more effective. In this text we'll be primarily interested

in how people acquire the rules of their language. Not all rules of grammar are acquired, however. Some facts about Language seem to be built into our brains, or ***innate***.

You now have enough information to answer General Problem Set 3

4.2 Innateness: Language as an Instinct

If you think about the other types of knowledge that are subconscious, you'll see that many[3] of them (for example, the ability to walk) are built directly into our brains – they are instincts. No one had to teach you to walk (despite what your parents might think!). Kids start walking on their own. Walking is an instinct. Probably the most controversial claim of Noam Chomsky's is that Language is also an instinct. Many parts of Language are built in, or ***innate***. Much of Language is an ability hard-wired into our brains by our genes.

Obviously, particular languages are not innate. It isn't the case that a child of Slovak parents growing up in North America who is never spoken to in Slovak, grows up speaking Slovak. They'll speak English (or whatever other language is spoken around them). So on the surface it seems crazy to claim that Language is an instinct. There are very good reasons to believe, however, that a human facility for Language (perhaps in the form of a "Language organ" in the brain) is innate. We call this facility ***Universal Grammar*** (or ***UG***).

4.3 The Logical Problem of Language Acquisition

What follows is a fairly technical proof of the idea that Language is at least plausibly construed as an innate, in-built system. If you aren't interested in this proof (and the problems with it), then you can reasonably skip ahead to section 4.4.

The argument in this section is that a productive system like the rules of Language probably have not been learned or acquired. Infinite systems are in principle, given certain assumptions, both unlearnable and unacquirable. Since we all have such an infinite system in our heads, and we shouldn't have been able to aquire it. So it follows that it is built in. The argument presented here is based on an unpublished paper by Alec Marantz, but is based on an argument dating back to at least Chomsky (1965).

First here's a sketch of the proof, which takes the classical form of an argument by modus ponens:

---

[3] but not all!

Premise (i): Syntax is a productive, recursive and infinite system
Premise (ii): Rule governed infinite systems are unlearnable.
Conclusion: Therefore syntax is an unlearnable system. Since we have it, it follows that at least parts of syntax are innate.

There are parts of this argument that are very controversial. In the challenge problem sets at the end of this chapter you are invited to think very critically about the form of this proof. Challenge Problem Set 3 considers the possibility that premise (i) is false (but hopefully, you will conclude that despite the argument given in the problem set, that the idea Language is productive and infinite is correct). Premise (ii) is more dubious, and is the topic of Challenge Problem Set 4. Here, in the main body of the text, I will give you the classic versions of the support for these premises, without criticizing them. You are invited to be skeptical and critical of them when you do the Challenge Problem sets.

Let's start with premise (i). Language is a productive system. That is, you can produce and understand sentences you have never heard before. For example, I can practically guarantee that you have never heard the following sentence:

12) The dancing chorus-line of elephants broke my television set.

The magic of syntax is that it can generate forms that have never been produced before. Another example of the productive quality lies in what is called ***recursion***. It is possible to utter a sentence like (13):

13) Rosie loves magazine ads.

It is also possible to put this sentence inside another sentence, like (14):

14) I think [Rosie loves magazine ads].

Similarly you can put this larger sentence inside of another one:

15) Drew believes [I think [Rosie loves magazine ads]].

and of course you can put this bigger sentence inside of another one:

16) Dana doubts that [Drew believes [I think [Rosie loves magazine ads]]].

and so on, and so on ad infinitum. It is always possible to embed a sentence inside of a larger one. This means that Language is a productive (probably infinite) system. There are no limits on what we can talk about. Other examples of the productivity of syntax can be seen in the fact that you can infinitely repeat adjectives (17) and you can infinitely add coordinated nouns to a noun phrase (18):

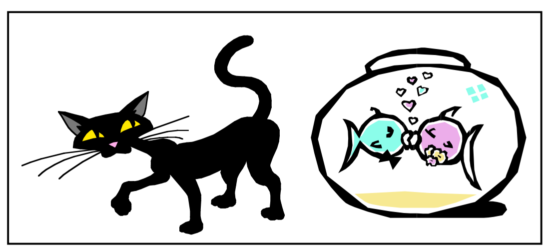
17) a) a very big peanut
b) a very very big peanut
c) a very very very big peanut
d) a very very very very big peanut
etc.

18) a) Dave left
b) Dave and Alina left
c) Dave, Dan and Alina left
d) Dave, Dan, Erin and Alina left
e) Dave, Dan, Erin, Jaime and Alina left
etc.

It follows that for every grammatical sentence of English, you can find a longer one (based on one of the rules of recursion, adjective repetition, or coordination) that is longer. This means that language is at least countably infinite. This premise is relatively uncontroversial (however, see the discussion in Challenge Problem Set 3).

Let's now turn to premise (ii): The idea that infinite systems are unlearnable. In order to make this more concrete, let's consider an algebraic treatment of a linguistic example. Imagine that the task of a child is to determine the rules by which her language is constructed. Further, let's simplify the task, and say a child simply has to match up situations in the real world with utterances she hears.[4] So upon hearing the utterance *the cat spots the kissing fishe*s, she identifies it with an appropriate situation in the context around her (as represented by the picture).

19) "the cat spots the kissing fishes" =



[4] The task is actually several magnitudes more difficult than this, as the child has to work out the phonology, etc., too, but for argument's sake, let's stick with this simplified example.

Her job, then, is to correctly match up the sentences with the situation.[5] More crucially she has to make sure that she does *not* match it up with all the other possible alternatives, such as the things going on around her (like her older brother kicking the furniture, or her mother making her breakfast, etc.). This matching of situations with expressions is a kind of mathematical relation (or function) that *maps* sentences onto a particular situation. Another way of putting it is that she has to figure out the rule(s) that decode(s) the meaning of the sentences. It turns out that this task is, at least very difficult if not impossible.

Let's make this even more abstract to get at the mathematics of the situation. Assign each sentence some number. This number will represent the input to the rule. Similarly we will assign each situation a number. The function (or rule) modeling language acquisition maps from the set of sentence numbers to the set of situation numbers. Now let's assume that the child has the following set of inputs and correctly matched situations (perhaps explicitly pointed out to her by her parents). The x value represents the sentences she hears. The y is the number correctly associated with the situation.

20)

Sentence (input)	*Situation* (output)
x	y
1	1
2	2
3	3
4	4
5	5

Given this input, what do you suppose that the output where $x = 6$ will be?

6	**?**

Most people will jump to the conclusion that the output will be 6 as well. That is, they assume that the function (the rule) mapping between inputs and outputs is $x = y$. But what if I were to tell you that in the hypothetical situation I envision here, the correct answer is situation number 126. The rule that generated the table in (20) is actually:

21) $[(x-5)*(x-4)*(x-3)*(x-2)*(x-1)] + x = y$

With this rule, all inputs equal to or less than 5 will give an output equal to the input, but for all inputs greater than 5, will give some large number.

---

[5] Note that this is the job of the child who is using universal grammar, not the job of UG itself.

When you hypothesized the rule was $x = y$, you didn't have all the crucial information; you only had part of the data. This seems to mean that if you hear only the first five pieces of data in our table then you won't get the rule, but if you learn the sixth you will figure it out. Is this necessarily the case? Unfortunately not: Even if you add a sixth line, you have no way of being sure that you have the right function until you have heard *all* the possible inputs. The important information might be in the sixth line, but it might also be in the 7,902,821,123,765th sentence that you hear. You have no way of knowing for sure if you have heard all the relevant data until you have heard them all. In an infinite system you can't hear them all, even if you were to hear 1 sentence every 10 seconds for your entire life. If we assume the average person lives to be about 75 years old, if they heard one new sentence every ten seconds, ignoring leap years and assuming they never sleep, they'd have only heard about 39,420,000 sentences over their lifetime. This is a much smaller number than infinity. Despite this poverty of input, by the age of 5 most children are fairly confident with their use of complicated syntax. Productive systems are (possibly) unlearnable, because you never have enough input to be sure you have all the relevant facts. This is called ***the logical problem of language acquisition***.

Generative grammar gets around this logical puzzle by claiming that the child acquiring English, Irish, or Yoruba has some help: a flexible blueprint to use in constructing her knowledge of language called Universal Grammar. Universal Grammar restricts the number of possible functions that map between situations and utterances, thus making language learnable.

You now have enough information to try Challenge Problem Sets 3 & 4

4.4 Other Arguments for UG

The evidence for UG doesn't rely on the logical problem alone, however. There are many other arguments that support the hypothesis that at least a certain amount of language is built in.

An argument that is directly related to the logical problem of language acquisition discussed above has to do with the fact that we know things about the grammar of our language that we couldn't possibly have learned. Start with the data in (20). A child might plausibly have heard sentences of these types (the underline represents the place where the question word *who* plausibly starts out – that is either as the object or subject of the verb *will question*):

22) a) Who do you think that Ciaran will question _____ first?
b) Who do you think Ciaran will question _____ first?
c) Who do you think _____ will question Seamus first?

The child has to draw a hypothesis about the distribution of the word *that* in English sentences. One conclusion consistent with this observed data is that the word *that* in English is optional. You can either have it or not. Unfortunately this conclusion is not accurate. Consider the fourth sentence in the paradigm in (22). This sentence is the same as (22c) but with a *that*:

d) *Who do you think that _____ will question Seamus first?

It appears as if *that* is only optional when the question word (*who* in this case) starts in object position (as in 22a and b) It is obligatorily absent when the question word starts in subject position (as in 22c and d) (don't worry about the details of this generalization). What is important to note is that *no one* has ever taught you that (22d) is ungrammatical. Nor could you have come to that conclusion on the basis of the data you've heard. The logical hypothesis on the basis of the data in (220a–c) predicts sentence (22d) to be grammatical. There is nothing in the input a child hears that would lead them to the conclusion that (22d) is ungrammatical, yet every English-speaking child knows it is. One solution to this conundrum is that we are born with the knowledge that sentences like (22d) are ungrammatical.[6] This kind of argument is often called the ***underdetermination of the data*** argument for UG.

Most parents raising a toddler will swear up and down that they are teaching their children to speak; that they actively engage in instructing their child in the proper form of the language. That overt instruction by parents plays any role in language development is easily falsified. The evidence from the experimental language acquisition literature is very clear: parents, despite their best intentions, do not, for the most part, correct ungrammatical utterances by their children. More generally, they correct the content rather

[6] The phenomenon in (22) is sometimes called the **that-*trace effect***. There is no disputing the fact that this phenomenon is not learnable. However, it is also a fact that it is not a universal property of all languages. For example, French and Irish don't seem to have the *that*-trace effect. Here is a challenge for those of you who like to do logic puzzles: If the *that*-trace effect is not learnable and thus must be biologically built in, how is it possible for a speaker of French or Irish to violate it? Think carefully about what kind of input a child might have to have in order to learn an "exception" to a built-in principle. This is a hard problem, but there is a solution. It may become clearer below when we discuss parameters.

than the form of their child's utterance (see for example the extensive discussion in Holzman 1997).

23) (from Marcus et al. 1992)
Adult: Where is that big piece of paper I gave you yesterday?
Child: Remember? I writed on it.
Adult: Oh that's right, don't you have any paper down here, buddy?

When a parent does try to correct a child's sentence structure, it is more often than not ignored by the child:

24) (from Pinker 1995: 281 – attributed to Martin Braine)
Child: Want other one spoon, Daddy
Adult: You mean, you want the other spoon.
Child: Yes, I want other one spoon, please Daddy.
Adult: Can you say "the other spoon"?
Child: Other … one … spoon
Adult: Say "other".
Child: other
Adult: "spoon"
Child: spoon
Adult: "other … spoon"
Child: other … spoon. Now give me other one spoon.

This humorous example is typical of parental attempts to "instruct" their children in language. When they do occur, they fail. However, children still acquire language in the face of a complete lack of instruction. Perhaps one of the most convincing explanations for this is UG. In the problem set part of this chapter, you are asked to consider other possible explanations and evaluate which are the most convincing.

There are also typological arguments for the existence of an innate language faculty. All the languages of the world share certain properties (for example they *all* have subjects and predicates – other examples will be seen throughout the rest of this book). These properties are called ***universals*** of Language. If we assume UG, then the explanation for these language universals is straightforward – they exist because all speakers of human languages share the same basic innate materials for building their language's grammar. In addition to sharing many similar characteristics, recent research into Language acquisition has begun to show that there is a certain amount of consistency cross-linguistically in the way children acquire Language. For example, children seem to go through the same stages and make the same kinds of mistakes when acquiring their language, no matter what their cultural background.

Statistical Probability or UG?

In looking at the logical problem of language acquisition you might be asking yourself "Ok, so maybe kids don't get all the data, but maybe they get enough to draw conclusions about what is the most likely structure of their grammar?" For example, we might conclude that a child learning English would observe the total absence of any sentences that have *that* followed by a trace (e.g., 22d), so after hearing some threshold of sentences they conclude that this sentence type is ungrammatical. This is a common objection to the hypothesis of UG. Unfortunately, this hypothesis can't explain why many sentence types that are extremely rare (to the point that they are probably never heard by children) are still judged as grammatical by the children. For example, English speakers rarely (if ever) produce sentences with seven embeddings (*John said that Mary thinks that Susan believes that Matt exclaimed that Marian claimed that Art said that Andrew wondered if Gwen had lost her pen*); yet speakers of English routinely agree these are acceptable. The actual speech of adult speakers is riddled with errors (due to all sorts of external factors: memory, slips of the tongue, tiredness, distraction, etc.). But children do not seem to assume that any of these errors, which they hear frequently, are part of the data that determines their grammars.

Finally, there are a number of biological arguments in favor of UG. As noted above, Language seems to be both human-specific and pervasive across the species. All humans, unless they have some kind of physical impairment, seem to have Language as we know it. This points towards it being a genetically endowed instinct. Additionally, research from neurolinguistics seems to point towards certain parts of the brain being linked to specific linguistic functions.

With very few exceptions, most linguists believe that some Language is innate. What is of controversy is how much is innate and whether the innateness is specific to Language, or follows from more general innate cognitive functions. We leave these questions unanswered here.

You now have enough information to try General Problem Set 4

4.5 Explaining Language Variation

The evidence for UG seems to be very strong. However, we are still left with the annoying problem that languages differ from one another. This problem is what makes the study of syntax so interesting. It is also not an unsolvable

one. One way in which languages differ is in terms of the words used in the language. These clearly have to be learned or memorized. Other differences between languages (such as the fact that basic English word order is subject-verb-object (SVO), but the order of an Irish sentence is verb-subject-object (VSO) and the order of a Turkish sentence is subject-object-verb (SOV)) must also be acquired. The explanation for this kind of fact will be explored in chapter 6. Foreshadowing slightly, we'll claim there that differences in the grammars of languages can be boiled down to the setting of certain innate ***parameters*** (or switches) that select among possible variants. Language variation thus reduces to learning the correct set of words and selecting from a predetermined set of options.

Oversimplifying slightly, most languages put the order of elements in a sentence in one of the following word orders:

25) a) Subject Verb Object (SVO) (e.g., English)
b) Subject Object Verb (SOV) (e.g., Turkish)
c) Verb Subject Object (VSO) (e.g., Irish)

A few languages use:

d) Verb Object Subject (VOS) (e.g., Malagasy)

No (or almost no)[7] languages use

e) Object Subject Verb (OSV)
f) Object Verb Subject (OVS)

Let us imagine that part of UG is a parameter that determines the basic word order. Four of the options (SVO, SOV, VSO, and VOS) are innately available as possible settings. Two of the possible word orders are not part of UG. The child who is acquiring English is innately biased towards one of the common orders, when she hears a sentence like "Mommy loves Kirsten," if the child knows the meaning of each of the words, then she might hypothesize two possible word orders for English: SVO and OVS. None of the others are consistent with the data. The child thus rejects all the other hypotheses. OVS is not allowed, since it isn't one of the innately available forms. This leaves SVO, which is the correct order for English. So children acquiring English will choose to set the word order parameter at the innately available SVO setting.

In his excellent book *The Atoms of Language*, Mark Baker inventories a set of possible parameters of a language variation within the UG hypothesis.

[7] This is a matter of some debate. Derbyshire (1985) has claimed that the language Hixkaryana has object initial order.

This is an excellent and highly accessible treatment of parameters. I strongly recommend this book.

You now have enough information to try General Problem Set 5 and Challenge Set 5

5. Choosing among Theories about Syntax

There is one last preliminary we have to touch on before actually doing some real syntax. In this book we are going to posit many hypotheses. Some of these we'll keep, others we'll revise, and still others we'll reject. How do we know what is a good hypothesis and what is a bad? Chomsky (1965) proposed that we can evaluate how good theories of syntax are, using what are called the ***levels of adequacy***. Chomsky claimed that there are three stages that a grammar (the collection of descriptive rules that constitute your theory) can attain in terms of adequacy.

If your theory only accounts for the data in a corpus (say a series of printed texts) and nothing more it is said to be an ***observationally adequate grammar***. Needless to say, this isn't much use if we are trying to account for the cognition of Language. As we discussed above, it doesn't tell us the whole picture. We also need to know what kinds of sentences are unacceptable, or ill-formed. A theory that accounts for both corpora and native speaker judgments about well-formedness is called a ***descriptively adequate grammar***. On the surface this may seem to be all we need. Chomsky, however, has claimed that we can go one step better. He points out that a theory that also accounts for how children acquire their language is the best. He calls this an ***explanatorily adequate grammar***. The simple theory of parameters might get this label. Generative grammar strives towards explanatorily adequate grammars.

You now have enough information to try General Problem Set 6

6. The Scientific Method and the Structure of this Textbook

Throughout this chapter I've emphasized the importance of the scientific method to the study of syntax. It's worth noting that we're not only going to apply this principle to small problems or specific rules, but we'll also apply it in a more global way. This principle is in part a guide to the way in which the rest of this book is structured.

In chapters 2–5 (the remainder of Part I of the book) we're going to develop an initial hypothesis about the way in which syntactic rules

are formed. These are the Phrase Structure Rules (PSRs). Chapters 2 and 3 examine the words these rules use, the form of the rules, and they structures they generate. Chapters 4 and 5 look at ways we can detail the structure of the trees formed by the PSRs.

In chapters 6–8 (Part 2 of the book), we examine some data that presents problems for the simple grammar presented in Part 1. When faced with more complicated data, we revise our hypotheses, and this is precisely what we do. We develop a special refined kind of PSR known as an X-bar rule. X-bar rules are still phrase structure rules, but they offer a more sophisticated way of looking at trees. This more sophisticated version also needs an additional constraint known as the "theta criterion" which is the focus of chapter 8.

In chapters 9–12 (Part 3) we consider even more data, and refine our hypothesis again. This time adding a new rule type: the transformation (we retain X-bar, but enrich it with transformations). Part 4 of the book (chapters 13–16) refines these proposals even further.

With each step we build upon our initial hypothesis, just as the scientific method tells us to. I've been teaching with this proposal-revision method theory construction for a couple of years now, and every now and then I hear the complaint from a student that we should just start with the final answer (i.e. the revised hypothesis found in the later chapters in the book). Why bother learning all this "other" "wrong" stuff? Why should we bother learning Phrase Structure Rules? Why don't we just jump straight into X-bar theory? Well, in principle, I could have constructed a book like that, but then you, the student, wouldn't understand *why* things are the way they are in the latter chapters. The theory would appear to be unmotivated, and you wouldn't understand what the technology actually does. By proposing a simple hypothesis early on in the initial chapters, and then refining and revising it, building new ideas onto old ones, you not only get an understanding of the motivations for and inner workings of our theoretical premises, but you get practice in working like a real linguist. Professional linguists, like all scientists, work from a set of simple hypotheses and revise them in light of predictions made by the hypotheses. The earlier versions of the theory aren't "wrong" so much as they need refinement and revision. These early versions represent the foundations out of which the rest of the theory has been built This is how science works.

7. SUMMARY

In this chapter, we've done very little syntax but talked a lot about the assumptions underlying the approach we're going to take to the study of sentence structure. The basic approach to syntax that we'll be using here is generative grammar; we've seen that this approach is scientific in that it uses the scientific method. It is descriptive and rule based. Further, it assumes that a certain amount of grammar is built in and the rest is acquired.

IDEAS, RULES, AND CONSTRAINTS INTRODUCED IN THIS CHAPTER

i) ***Syntax***: The level of linguistic organization that mediates between sounds and meaning, where words are organized into phrases and sentences.

ii) ***Language*** *(capital L)*: The psychological ability of humans to produce and understand a particular language. Also called the ***Human Language Capacity*** or ***i-Language***. This is the object of study in this book.

iii) ***language*** *(lower-case l)*: A language like English or French. These are the particular instances of the human Language. The data source we use to examine Language is language. Also called ***e-language***.

iv) ***Generative Grammar***: A theory of linguistics in which grammar is viewed as a cognitive faculty. Language is generated by a set of rules or procedures. The version of generative grammar we are looking at here is primarily the ***Principles and Parameters approach*** (P&P) touching occasionally on ***Minimalism.***

v) ***The Scientific Method***: Observe some data, make generalizations about that data, draw a hypothesis, test the hypothesis against more data.

vi) ***Falsifiable Prediction***: To prove that a hypothesis correct you have to look for the data that would prove it *wrong*. The prediction that might prove a hypothesis wrong is said to be falsifiable.

vii) ***Grammar***: Not what you learned in school. This is the set of rules that generate a language.

viii) ***Prescriptive Grammar***: The grammar rules as taught by so called "language experts." These rules, often inaccurate descriptively,

prescribe how people should talk/write, rather than describe what they actually do.

ix) ***Descriptive Grammar***: A scientific grammar that describes, rather than prescribes, how people talk/write.

x) ***Anaphor***: A word that ends in *-self* or *-selves* (a better definition will be given in chapter 5).

xi) ***Antecedent***: The noun an anaphor refers to.

xii) ***Asterisk***: * used to mark syntactically ill-formed (unacceptable or ungrammatical) sentences. The hash mark, pound, or number sign (#) is used to mark semantically strange, but syntactically well-formed, sentences.

xiii) ***Gender (Grammatical)***: Masculine vs. Feminine vs. Neuter. Does not have to be identical to the actual sex of the referent. For example, a dog might be female, but we can refer to it with the neuter pronoun *it*. Similarly, boats don't have a sex, but are grammatically feminine.

xiv) ***Number***: The quantity of individuals or things described by a noun. English distinguishes singular (e.g., *a cat*) from plural (e.g., *the cats*). Other languages have more or less complicated number systems.

xv) ***Person***: The perspective of the participants in the conversation. The speaker or speakers (*I, me, we, us*) are called first person. The listener(s) (*you*), are called the second person. Anyone else (those not involved in the conversation) (*he, him, she, her, it, they, them*) is referred to as the third person.

xvi) ***Case***: The form a noun takes depending upon its position in the sentence. We discuss this more in chapter 10.

xvii) ***Nominative***: The form of a noun in subject position (*I, you, he, she, it, we, they*).

xviii) ***Accusative***: The form of a noun in object position (*me, you, him, her, it, us, them*).

xix) ***Corpus (pl. Corpora)***: A collection of real-world language data.

xx) ***Native Speaker Judgments (intuitions)***: Information about the subconscious knowledge of a language. This information is tapped by means of the grammaticality judgment task.

xxi) ***Semantic Judgment***: A judgment about the meaning of a sentence, often relying on our knowledge of the context in which the sentence was uttered.

xxii) ***Syntactic Judgment***: A judgment about the form or structure of a sentence.

xxiii) ***Learning***: The gathering of conscious knowledge (like linguistics or chemistry).

xxiv) ***Acquisition***: The gathering of subconscious information (like language).

xxv) ***Innate***: Hard-wired or built in, an instinct.

xxvi) ***Recursion***: The ability to embed structures iteratively inside one another. Allows us to produce sentences we've never heard before.

xxvii) ***Universal Grammar (UG)***: The innate (or instinctual) part of each language's grammar.

xxviii) ***The Logical Problem of Language Acquisition***: The proof that an infinite system like human language cannot be learned on the basis of observed data – an argument for UG.

xxix) ***Underdetermination of the Data***: The idea that we know things about our language that we could not have possibly learned – an argument for UG.

xxx) ***Universal***: A property found in all the languages of the world.

xxxi) ***Observationally Adequate Grammar***: A grammar that accounts for observed real-world data (such as corpora).

xxxii) ***Descriptively Adequate Grammar***: A grammar that accounts for observed real-world data and native speaker judgments.

xxxiii) ***Explanatorily Adequate Grammar***: A grammar that accounts for observed real-world data and native speaker judgments and offers an explanation for the facts of language acquisition.

FURTHER READING

Baker, Mark (2001) *The Atoms of Language: The Mind's Hidden Rules of Grammar*. New York: Basic Books.

Barsky, Robert (1997) *Noam Chomsky: A Life of Dissent*. Cambridge: MIT Press.
Chomsky, Noam (1965) *Aspects of the Theory of Syntax*. Cambridge: MIT Press.
Jackendoff, Ray (1993) *Patterns in the Mind.* London: Harvester-Wheatsheaf.
Pinker, Steven (1995) *The Language Instinct*. New York: Harper Perennial.
Sampson, Geoffrey (1997) *Educating Eve: The Language Instinct Debate*. London: Cassell.
Uriagereka, Juan (1998) *Rhyme and Reason: An Introduction to Minimalist Syntax*. Cambridge: MIT Press.

General Problem Sets

1. Prescriptive Rules

[Creative and Critical Thinking; Basic]

In the text above, we argued that descriptive rules are the primary focus of syntactic theory. This doesn't mean that prescriptive rules don't have their uses. What are these uses? Why do we maintain prescriptive rules in our society?

2. Judgments

[Application of Skills; Intermediate]

All of the following sentences have been claimed to be ungrammatical or unacceptable by someone at some time. For each sentence, indicate whether this unacceptability is

i) a prescriptive or a descriptive judgment, and
ii) for all descriptive judgments indicate whether the ungrammaticality has to do with syntax or semantics (or both).

One- or two-word answers are appropriate. If you are not a native speaker of English, enlist the help of someone who is. If you are not familiar with the *prescriptive* rules of English grammar, you may want to consult a writing guide or English grammar or look at Pinker's *The Language Instinct*.

a) Who did you see in Las Vegas?
b) You are taller than me.
c) My red is refrigerator.
d) Who do you think that saw Bill?
e) Hopefully, we'll make it through the winter without snow.
f) My friends wanted to quickly leave the party.
g) Bunnies carrots eat.
h) John's sister is not his sibling.

3. Learning vs. Acquisition

[Creative and Critical Thinking; Basic]

We have distinguished between learning and acquiring knowledge. Learning is conscious, acquisition is automatic and subconscious. (Note that acquired things are *not* necessarily innate. They are just subconsciously obtained.) Other than language are there other things we acquire? What other things do we learn? What about walking? Or reading? Or sexual identity? An important point in answering this question is to talk about what kind of evidence is necessary to distinguish between learning and acquisition.

4. Universals

[Creative and Critical Thinking; Intermediate]

Pretend for a moment that you don't believe Chomsky and that you don't believe in the innateness of syntax (but only *pretend*!). How might you account for the existence of universals (see definition above) across languages?

5. Innateness

[Creative and Critical Thinking; Intermediate]

We argued that some amount of syntax is innate (inborn). Can you think of an argument that might be raised against innateness? (It doesn't have to be an argument that works, just a plausible one.) Alternately, could you come up with a hypothetical experiment that could *disprove* innateness? What would such an experiment have to show? Remember that cross-linguistic variation (differences between languages) is *not* an argument against innateness or UG, because UG contains parameters that allow minute variations.

6. Levels of Adequacy

[Application of Skills; Basic]

Below, you'll find the description of several different linguists' work. Attribute a level of adequacy to them (state whether the grammars they developed are observationally adequate, descriptively adequate, or explanatorily adequate). Explain *why* you assigned the level of adequacy that you did.

a) Juan Martínez has been working with speakers of Chicano English in the barrios of Los Angeles. He has been looking both at corpora (rap music, recorded snatches of speech) and working with adult native speakers.
b) Fredrike Schwarz has been looking at the structure of sentences in eleventh-century Welsh poems. She has been working at the national archives of Wales in Cardiff.
c) Boris Dimitrov has been working with adults and corpora on the formation of questions in Rhodopian Bulgarian. He is also conducting a longitudinal study of some two-year-old children learning the language to test his hypotheses.

CHALLENGE PROBLEM SETS

Challenge Problem Sets are special exercises that either challenge the presentation of the main text or offer significant enrichment. Students are encouraged to complete the other problem sets before trying the Challenge Sets. Challenge Sets can vary in level from interesting puzzles to downright impossible conundrums. Try your best!

CHALLENGE PROBLEM SET 1: ANAPHORA

[Creative and Critical Thinking and Data Analysis; Challenge]

In this chapter, as an example of the scientific method, we looked at the distribution of anaphora (nouns like *himself*, *herself*, etc.). We came to the following conclusion about their distribution:

> An anaphor must agree in person, gender, and number with its antecedent.

However, there is much more to say about the distribution of these nouns (in fact, chapter 5 of this book is entirely devoted to the question).

Part 1: Consider the data below. Can you make an addition to the above statement that explains the distribution of anaphors and antecedents in the very limited data below?

a) Geordi sang to himself.
b) *Himself sang to Geordi.
c) Betsy loves herself in blue leather.
d) *Blue leather shows herself that Betsy is pretty.

Part 2: Now consider the following sentences:[8]

e) Everyone should be able to defend himself/herself/themselves.
f) I hope nobody will hurt themselves/himself/?herself.

Do these sentences obey your revised generalization? Why or why not? Is there something special about the antecedents that forces an exception here, or can you modify your generalization to fit these cases?

CHALLENGE PROBLEM SET 2: YOURSELF

[Creative and Critical Thinking; Challenge]

In the main body of the text we claimed that all anaphors need an antecedent. Consider the following acceptable sentence. This kind of sentence is called an "imperative" and is used to give orders.

a) Don't hit yourself!

[8] Thanks to Ahmad Lotfi for suggesting this part of the question.

Part 1: Are all anaphors allowed in sentences like (a)? Which ones are allowed there, and which ones aren't.

Part 2: Where is the antecedent for yourself? Is this a counter-example to our rule? Why is this rule an exception? It is easy to add a stipulation to our rule; but we'd rather have an explanatory rule. What is special about the sentence in (a)?

CHALLENGE PROBLEM SET 3: IS LANGUAGE REALLY INFINITE?
[Creative and Critical Thinking; Extra Challenge]

[***Note to instructors***: this question requires some background either in formal logic or mathematical proofs.]

In the text, it was claimed that because language is recursive, it follows that it is infinite. (This was premise (i) of the discussion in section 4.3). The idea is straightforward and at least intuitively correct: if you have some well-formed sentence, and you have a rule that can embed it inside another structure; then you can also take this new structure and embed it inside another and so on and so on. Intuitively this leads to an infinitely large number of possible sentences. Pullum and Scholz (2005) have shown that one formal version of this intuitive idea is either circular or a contradiction.

Here is the structure of the traditional argument (paraphrased and simplified from the version in Pullum and Scholz). This proof is cast in such a way that the way we count the number of sentences is by comparing the number of words in the sentence. If for *any* (extremely high) number of words, we can find a longer sentence, then we know the set is infinite. First some terminology:

- *Terminology:* call the set of well-formed sentences *E*. If a sentence *x* is an element of this set we write *E(x)*.
- *Terminology:* let us refer to the length of a sentence by counting the number of words in it. The number of words in a sentence is expressed by the variable *n.* There is a special measurement operation (function) which counts the number of words, this is called μ. If the sentence called *x* has 4 words in it then we say μ(x) = 4.

Next the formal argument:

Premise 1: There is at least one well-formed sentence that has more than zero words in it.

$$\exists x[E(x) \ \& \ \mu(x) > 0]$$

Premise 2: There is an operation in the PSRs such that any sentence may be embedded in another with more words in it. That means for any sentence in the language, there is another longer sentence. (If some expression has the length *n*, then some other well-formed sentence has a size greater than *n*).

$$\forall n \ [\exists x[E(x) \ \& \ \mu(x) = n]] \rightarrow [\exists y[E(y) \ \& \ \mu(y) > n]]$$

Conclusion: Therefore for every positive integer *n*, there are well-formed sentences with a length longer than *n* (i.e., the set of well-formed English expressions is at least countably infinite):

$$\therefore \forall n\ [\exists y[E(y)\ \&\ \mu(y) > n]]$$

Pullum and Scholz claim that the problem with this argument lies with the nature of the set E. Sets come of two kinds: there are finite sets which have a fixed number of elements (e.g. the set {a, b, c, d} has 4 and exactly 4 members). There are also infinite sets, which have an endless possible number of members (e.g., the set {a, b, c, … } has an infinite number of elements).

Question 1: Assume that E, the set of well-formed sentences, is finite. This is a contradiction of one of the two premises given above. Which one? Why is it a contradiction?

Question 2: Assume that E, the set of well-formed sentences, is infinite. This leads to a circularity in the argument. What is the circularity (i.e., why is the proof circular)?

Question 3: If the logical argument is either contradictory or circular what does that make of our claim that the number of sentences possible in a language is infinite? Is it totally wrong? What does the proof given immediately above really prove?

Question 4: Given that E can be neither a finite nor an infinite set, is there any way we might recast the premises, terminology, or conclusion in order not to have a circular argument and capture the intuitive insight of the claim? Explain how we might do this or why it's impossible. Try to be creative. There is no "right" answer to this question. Hint: one might try a proof that proves that a subset of the sentences of English is infinite (and by definition the entire set of sentences in English is infinite) or one might try a proof by contradiction.

Important notes:

1) Your answers can be given in English prose, you do not need to give a formal mathematical answer.
2) Do not try to look up the answer in the papers cited above. That's just cheating! Try to work out the answers for yourself.

CHALLENGE PROBLEM SET 4: ARE INFINITE SYSTEMS REALLY UNLEARNABLE?

[Creative and Critical Thinking; Challenge]

In section 4.3, you saw the claim that if language is an infinite system then it must be unlearnable. In this problem set, you should aim a critical eye at the premise that infinite systems can't be learned on the basis of the data you hear.

While given the extreme view in section 4.3 is logically true, consider the following alternative possibilities:

a) We as humans have some kind of "cut off mechanism" that stops considering new data after we've heard some threshold number of examples. If we don't hear the crucial example after some period of time we simply assume it doesn't exist. Rules simply can't exist that require access to sentence types so rare that you don't hear them before the cut off point.

b) We are purely statistical engines. Rare sentences types are simply ignored as "statistical noise." We consider only those sentences that are frequent in the input when constructing our rules.

c) Child-directed speech (motherese) is specially designed to give you precisely the kinds of data you need to construct your rule system. The child listens for very specific "triggers" or "cues" in the parental input in order to determine the rules.

Question 1: To what extent are (a), (b) or (c) compatible with the hypothesis of Universal Grammar. If (a), (b) or (c) turned out to be true, would this mean that there was no innate grammar? Explain your answer.

Question 2: How might you experimentally or observationally distinguish between (a), (b), (c) and the infinite input hypothesis of 4.3? What kinds of evidence would you need to tell them apart?

Question 3: When people speak, they make errors. (They switch words around, they mispronounce things, they use the wrong word, they stop mid-sentence without completing what they are saying etc.) Nevertheless children seem to be able to ignore these errors and still come up with the right set of rules. Is this fact compatible with any of the infinite hypothesis, (a), (b), or (c)?

Challenge Problem Set 5: Learning Parameters: Pro Drop

[Critical Thinking, Data Analysis; Challenge]

Background: Among the Indo-European languages there are two large groups of languages that pattern differently with respect to whether they require a pronoun (like he, she, it) in the subject position, or whether such pronouns can be "dropped". For example, in both English and French, pronouns are required. Sentences without them are usually ungrammatical:

a) He left
b) *Left
c) Il est parti (French)
he is gone
"he left"
d) *est parti (French)

In languages, such as Spanish and Italian, however, such pronouns are routinely omitted (1S = first person, singular):

e) Io telephono (Italian)
I called.1S
"I called (phoned)"

f) telephono
called.1S
"Called"

Question 1: Now imagine that you are a small child learning a language. What kind of data would you need to know in order to tell if your language was "pro drop" or not? (Hint. Does the English child hear sentences both with and without subjects? Does the Italian child? Are they listening for sentences with subjects or without them?

Question 2: Assume that one of the two possible settings for this parameter (either your language is pro-drop or it is not) is the "default" setting. This default setting is the version of the parameter one gets if one doesn't hear the right kind of input. Which of the two possibilities is the default?

Question 3: English has imperative constructions such as:

g) Leave now!

Why doesn't the English child assume on the basis of such sentences that English is pro-drop?